INFORMATION AND CONTROL **6**, 297–300 (1963)

Entropy of Tamil Prose

GIFT SIROMONEY

Madras Christian College, Tambaram, India

The proportions of the different letters of the alphabet in Tamil prose are estimated from a large sample and an optimum code is constructed. The prose is compared with the Tamil poetry of different periods and the one-gram entropy of prose is significantly different from the entropies of the poetical works considered. An estimate is made, experimentally, of the entropy of Tamil prose.

Let $p_1, p_2, \cdots, p_n$ be the proportions of the different letters of the alphabet. The one-gram entropy (Shannon, 1948) is given by

$$H_1 = -\sum_1^n p_r \operatorname{ld} p_r,$$

where "ld" stands for the logarithm to the base two. If all the p_r are equal, then the value of H_1 is denoted by $H_0 = \operatorname{ld} n$.

In modern Tamil prose, 12 vowels, 18 consonants, 216 vowel-consonants and one auxiliary "Aitham" are used. In our discussion, each vowel-consonant will be considered as two letters—a consonant followed by a vowel. "Aitham" very seldom occurs in modern prose and it will not be considered; the vowels and consonants alone will be considered to make up the Tamil alphabet of 30 letters.

A large sample of over 20,000 letters was taken from the prose works published in Madras State during 1946–57, using random sampling methods.

$$H_1 = 4.34 \text{ bits} \quad \text{and} \quad H_0 = 4.91 \text{ bits}$$

For English prose (26-letter alphabet) the corresponding values are 4.14 and 4.70 bits respectively (Shannon, 1951).

Samples were taken from Tamil poetry from six different works belonging to periods ranging from the beginning of the Christian era to the modern period. As the dates of these works cannot be fixed with any certainty, the (accepted) order in which they were written is fol-

lowed. Tholkappiam (Porul athikaram) is the oldest and Bharathi's is the most recent. The values of the unbiased estimates of H_1 and the size of the sample used are tabulated (Table I) together with the corresponding standard deviations of the estimate of H_1. H_1 is asymptotically normally distributed and the unbiased estimate of H_1 and the standard deviation are obtained from the following formulas (Basharin, 1959).

$$E(\hat{H}_1) = H_1 - \frac{n-1}{2N} \operatorname{ld} e + O\left(\frac{1}{N^2}\right)$$

$$D(\hat{H}_1) = \frac{1}{N}\left[\sum_1^n p_r \operatorname{ld}^2 p_r - H_1^2\right] + O\left(\frac{1}{N^2}\right)$$

where N is the size of the sample.

H_1 is not significantly different between Thiru Kural and Silappathikaram of the early period and Bharathi's works of the modern period. Tholkappiam and Yuthakandam have practically the same value. The values of H_1 for Yuthakandam and Utharakandam are not significantly different and these two *Kandams* of Kamba Ramayanam are commonly accepted as written by two different authors, the author of the first being Kamban. A χ^2-test (Herdan, 1956) shows that the proportions of letters from these works are significantly different. Therefore, for the purposes of testing passages of disputed authorship the χ^2-test is a more useful tool than the characteristic H_1. The value of H_1 for Tamil prose is significantly different from those of all the poetic works considered. The letter "l-," which has a sound which is difficult to pronounce and peculiar to Tamil, has its highest value for Tholkappiam and the lowest for modern prose. Another peculiar sound represented by "t" shows the same type of downward trend towards the modern period.

H_1 is the average number of binary digits required, per letter of Tamil, if the language is encoded with 100% efficiency, on the first assumption that the occurrence of a letter in Tamil is independent of the preceding letters. Following Huffman (1952), an optimum binary code is constructed and tabulated. The average number of binary digits per symbol is 4.44 compared to 4.34, the value of H_1. Therefore the efficiency (Reza, 1961) of coding is 98%. Other codes which are not optimum, may be constructed, using the given values of the p's.

TABLE I

RELATIVE FREQUENCIES OF LETTERS

Letters	Tholka-ppiam	Thiru Kural	Sila-ppat-hikaram	Kam-ban's Yutha Kan-dam	Kam-ban's Uthara Kan-dam	Bhar-athi	Modern prose	Huffman code
a	139	126	142	145	156	129	150	101
a:	29	60	35	43	47	43	47	0100
i	83	64	68	69	71	75	78	1101
i:	3	7	7	5	6	7	4	01111010
u	87	74	73	72	58	70	77	1100
u:	5	5	6	6	6	3	4	011110111
e	19	24	22	23	20	22	17	011111
e:	14	11	11	11	12	20	13	010101
ai	29	30	29	23	28	32	27	01100
o	14	13	13	12	9	13	10	000100
o:	9	9	12	11	11	12	7	0111100
au	0	0	0	0	0	0	0	0111101100
k	56	71	67	58	63	49	79	1110
ng	9	7	11	9	9	5	6	0101000
c	16	21	18	17	19	24	22	00011
nj	5	4	3	2	4	3	1	0111101101
t:	25	25	32	27	25	25	36	10001
n:	15	20	15	16	16	16	11	000101
th	61	57	63	68	65	73	71	1001
nh	20	26	24	22	22	27	21	000000
p	49	42	39	33	33	40	44	0010
m	47	45	44	44	48	57	42	11111
y	38	34	38	36	35	36	29	01101
r	39	39	46	47	40	38	45	0011
l	33	35	33	32	34	36	31	01110
v	44	35	38	39	41	42	35	10000
l-	16	8	12	9	9	10	7	010100
l:	12	12	12	16	12	16	21	000001
t	41	43	36	36	32	24	27	01011
n	44	53	51	68	68	53	38	11110
	1001	1000	1000	999	999	1000	1000	
Total size of sample	12430	6355	4165	11855	16283	5056	22855	
UBE of H_1	4.4122	4.4621	4.4506	4.4150	4.3981	4.4661	4.3435	
S.D. of H_1	0.0094	0.0121	0.0158	0.0096	0.0079	0.0137	0.0073	

Our first assumption that the occurrences of the letters are independent can be improved by assuming that the occurrence of a letter is dependent on the $(n - 1)$ preceding letters. The corresponding n-gram entropy is denoted by H_n, and its limiting value, when n is large, is denoted by H. An estimate of H is made, using Shannon's (1951) experimental methods. Passages of lengths about of 30 letters were chosen from normal prose and the subject was asked to guess each letter. If the guess was correct, the subject was told so, and if the guess was wrong, the correct letter was given. The same passages were tried separately on two subjects S_1 and S_2, and S_1 guessed 320 letters and S_2, 317 out of 583 letters, giving 55% as the proportion of letters guessed correctly. S_1 and S_2 were second year university students. Following Brillouin (1956), the estimate of H is 2.51 bits and this should be considered as an upper limit. Therefore, it is possible to reduce drastically the length of a given message in Tamil, if proper encoding is used.

ACKNOWLEDGMENT

The author wishes to thank Dr. W. F. Kibble of Heriot-Watt College, Edinburgh and late Dr. R. P. Sethu Pillai, Professor of Tamil, University of Madras, for their valuable help at the initial stages of the work.

REFERENCES

BRILLOUIN, L. (1956), "Science and Information Theory," p. 25. Academic Press, New York.

BASHARIN, G. P. (1959), *Teoriya Veroyatnostei iee Primeneniya* **4**, 361.

HERDAN, G. (1956), "Language as Choice and Chance," p. 88. Noordhoff, Groningen.

HUFFMAN, D. A. (1952), *Proc. Inst. Radio Engrs.* **40**, 1098.

REZA, F. M. (1961), "An Introduction to Information Theory," p. 133. McGraw-Hill, New York.

SHANNON, C. E. (1948), *Bell System Tech. J.* **27**, 379.

SHANNON, C. E. (1951), *Bell System Tech. J.* **30**, 50.